CS-GY 6763: Lecture 13
Introduction to Spectral Sparsification

NYU Tandon School of Engineering, Prof. Rajesh Jayaram

Announcements:

- Final Exam next week during class time.
- Will have full **2 hours** for the exam.

BACK TO SUBSPACE EMBEDDINGS

Theorem (Subspace Embedding)

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ *be a matrix. If* $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$ *is chosen from any distribution* $\mathcal{D}$ *satisfying the Distributional JL Lemma, then with probability* $1 - \delta$,

$$(1-\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2$$

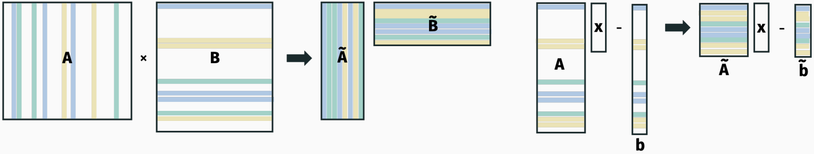
for *all* $\mathbf{x} \in \mathbb{R}^d$, *as long as* $m = O\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$.

Implies regression result, and more.

Example: The any singular value $\tilde{\sigma}_i$ of $\mathbf{\Pi}\mathbf{A}$ is a $(1 \pm \epsilon)$ approximation to the true singular value σ_i of $\mathbf{B}$.

SUBSAMPLING METHODS

Important Goal: Replace random projection methods with random sampling methods. Prove that for essentially all problems of interest, can obtain same asymptotic runtimes.

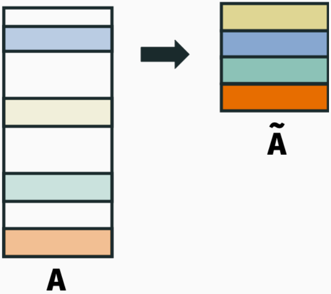


Sampling has the added benefit of preserving matrix sparsity or structure, and can be applied in a wider variety of settings where random projections are too expensive.

SUBSAMPLING METHODS

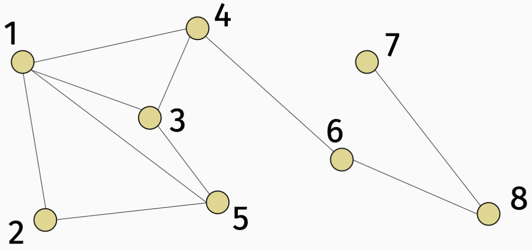
Goal: Can we use sampling to obtain subspace embeddings? I.e. for a given $\mathbf{A}$ find $\tilde{\mathbf{A}}$ whose rows are a (weighted) subset of rows in $\mathbf{A}$ and:

$$(1-\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \leq \|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2.$$



EXAMPLE WHERE STRUCTURE MATTERS

Let $\mathbf{B}$ be the edge-vertex incidence matrix of a graph G with vertex set V, $|V| = d$. Recall that $\mathbf{B}^T\mathbf{B} = \mathbf{L}$.

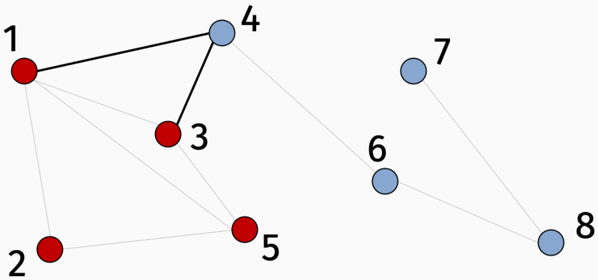


$$\mathbf{B} = \begin{bmatrix} +1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ +1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ +1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ +1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & +1 & 0 & -1 & 0 & 0 & 0 \\ & & & \vdots & & & & \\ 0 & 0 & 0 & +1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & +1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 \end{bmatrix}$$

Recall that if $\mathbf{x} \in \{-1, 1\}^n$ is the cut indicator vector for a cut S in the graph, then $\frac{1}{4}\|\mathbf{B}\mathbf{x}\|_2^2 = \text{cut}(S, V \setminus S)$.

LINEAR ALGEBRAIC VIEW OF CUTS

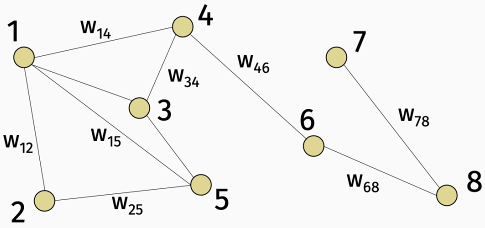
$$\mathbf{x} = [1, 1, 1, -1, 1, -1, -1, -1]$$

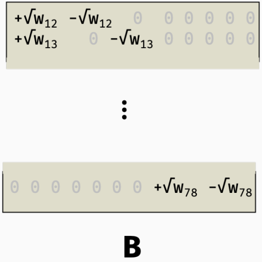


$\mathbf{x} \in \{-1, 1\}^d$ is the cut indicator vector for a cut S in the graph, then $\frac{1}{4}\|\mathbf{Bx}\|_2^2 = \text{cut}(S, V \setminus S)$

WEIGHTED CUTS

Extends to weighted graphs, as long as square root of weights is included in $\mathbf{B}$. Still have the $\mathbf{B}^T\mathbf{B} = \mathbf{L}$.

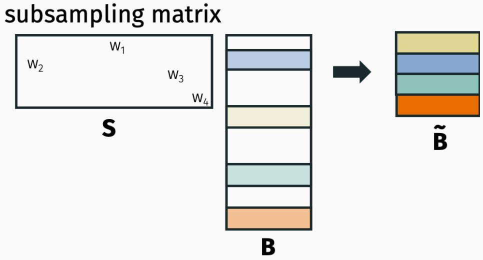




And still have that if $\mathbf{x} \in \{-1, 1\}^d$ is the cut indicator vector for a cut S in the graph, then $\frac{1}{4}\|\mathbf{Bx}\|_2^2 = \text{cut}(S, V \setminus S)$.

SPECTRAL SPARSIFICATION

Goal: Approximate $\mathbf{B}$ by a weighted subsample. I.e. by $\tilde{\mathbf{B}}$ with $m \ll |E|$ rows, each of which is a scaled copy of a row from $\mathbf{B}$.

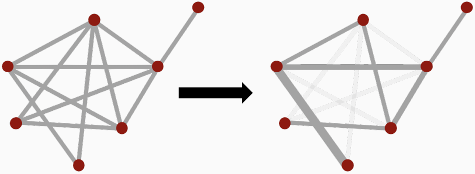


Natural goal: $\tilde{\mathbf{B}}$ is a subspace embedding for $\mathbf{B}$. In other words, $\tilde{\mathbf{B}}$ has $\approx O(d)$ rows and for all $\mathbf{x}$,

$$(1-\epsilon)\|\mathbf{B}\mathbf{x}\|_2^2 \leq \|\tilde{\mathbf{B}}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{B}\mathbf{x}\|_2^2.$$

HISTORY SPECTRAL SPARSIFICATION

$\tilde{\mathbf{B}}$ is itself an edge-vertex incidence matrix for some sparser graph $\tilde{G}$, which preserves many properties about G! $\tilde{G}$ is called a spectral sparsifier for G.



For example, we have that for any set S,

$$(1-\epsilon)\,\text{cut}_G(S, V \setminus S) \leq \text{cut}_{\tilde{G}}(S, V \setminus S) \leq (1+\epsilon)\,\text{cut}_G(S, V \setminus S).$$

So $\tilde{G}$ can be used in place of G in solving e.g. max/min cut problems, balanced cut problems, etc.

In contrast $\mathbf{\Pi B}$ would look nothing like an edge-vertex incidence matrix if $\mathbf{\Pi}$ is a JL matrix.

HISTORY OF SPECTRAL SPARSIFICATION

Spectral sparsifiers were introduced in 2004 by Spielman and Teng in an influential paper on faster algorithms for solving Laplacian linear systems.

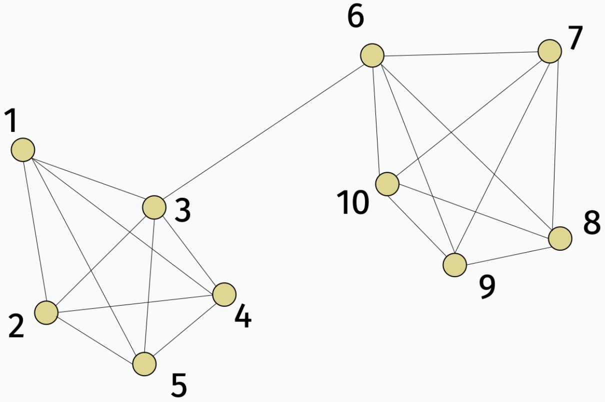
- Generalize the cut sparsifiers of Benczur, Karger '96.
- Further developed in work by Spielman, Srivastava + Batson, '08.
- Have had huge influence in algorithms, and other areas of mathematics – this line of work lead to the 2013 resolution of the Kadison-Singer problem in functional analysis by Marcus, Spielman, Srivastava.

Rest of class: Learn about an important random sampling algorithm for constructing spectral sparsifiers, and subspace embeddings for matrices more generally.

NATURAL FIRST ATTEMPT

Goal: Find $\tilde{\mathbf{A}}$ such that $\|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 = (1 \pm \epsilon)\|\mathbf{A}\mathbf{x}\|_2^2$ for all $\mathbf{x}$.

Possible Approach: Construct $\tilde{\mathbf{A}}$ by uniformly sampling rows from $\mathbf{A}$.



Can check that this approach fails even for the special case of a graph vertex-edge incidence matrix.

IMPORTANCE SAMPLING FRAMEWORK

Key idea: Importance sampling. Select some rows with higher probability.

Suppose $\mathbf{A}$ has n rows $\mathbf{a}_1 \ldots, \mathbf{a}_n$. Let $p_1, \ldots, p_n \in [0, 1]$ be sampling probabilities. Construct $\tilde{\mathbf{A}}$ as follows:

- For $i = 1, \ldots, n$
  - Select $\mathbf{a}_i$ with probability p_i.
  - If $\mathbf{a}_i$ is selected, add the scaled row $\frac{1}{\sqrt{p_i}}\mathbf{a}_i$ to $\tilde{A}$.

Remember, ultimately want that $\|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 = (1 \pm \epsilon)\|\mathbf{A}\mathbf{x}\|_2^2$ for all $\mathbf{x}$.

Claim 1: $\mathbb{E}[\|\tilde{\mathbf{A}}\mathbf{x}\|_2^2] = \|\mathbf{A}\mathbf{x}\|_2^2$.

Claim 2: Expected number of rows in $\tilde{\mathbf{A}}$ is $\sum_{i=1}^{n} p_i$.

LECTURE OUTLINE

How should we choose the probabilities $p_1, \ldots, p_n$?

1. Introduce the idea of row **leverage scores**.
2. Motivate why these scores make for good sampling probabilities.
3. Prove that sampling with probabilities proportional to these scores yields a subspace embedding (or a spectral sparsifier) with a near optimal number of rows.

MAIN RESULT

Let $\mathbf{A} = \mathbf{U\Sigma V}^T$ be the SVD of $\mathbf{A} \in \mathbb{R}^{n \times d}$. We define the **statistical leverage score** τ_i of row $\mathbf{A}_i$ as:

$$\tau_i = \|\mathbf{U}_i\|_2^2$$

i.e., τ_i is the norm of the *i*-th row of the left singular vector matrix $\mathbf{U} \in \mathbb{R}^{n \times d}$.

- We will show that τ_i is a natural importance measure for each row in $\mathbf{A}$.

MAIN RESULT

Let $\mathbf{A} = \mathbf{U\Sigma V}^T$ be the SVD of $\mathbf{A} \in \mathbb{R}^{n \times d}$. We define the **statistical leverage score** τ_i of row $\mathbf{A}_i$ as:

$$\tau_i = \|\mathbf{U}_i\|_2^2$$

i.e., τ_i is the norm of the *i*-th row of the left singular vector matrix $\mathbf{U} \in \mathbb{R}^{n \times d}$.

- We will show that τ_i is a natural importance measure for each row in $\mathbf{A}$.

Fact: We have that $\tau_i \in [0, 1]$ for all $i \in [n]$, and $\sum_{i=1}^n \tau_i = d$ if $\mathbf{A}$ is rank d.

- Follows from orthonormality of columns of $\mathbf{U}$

MAIN RESULT

For $i = 1, \ldots, n$,

$$\tau_i = \|\mathbf{U}_i\|_2^2$$

Theorem (Subspace Embedding from Subsampling)

For each i, and fixed constant c, let $p_i = \min\left(1, \frac{c \log d}{\epsilon^2} \cdot \tau_i\right)$. Let $\tilde{\mathbf{A}}$ have rows sampled from $\mathbf{A}$ with probabilities $p_1, \ldots, p_n$. With probability $9/10$, for all $x \in \mathbb{R}^d$.

$$(1-\epsilon)\|\mathbf{Ax}\|_2^2 \leq \|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{Ax}\|_2^2,$$

and $\tilde{\mathbf{A}}$ has $O(d \log d/\epsilon^2)$ rows in expectation.

VECTOR SAMPLING

How should we choose the probabilities $p_1, \ldots, p_n$?

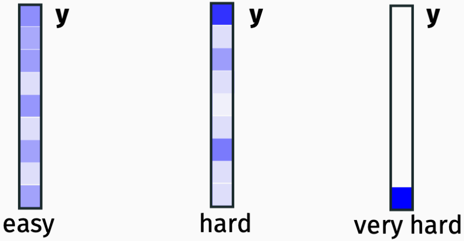
As usual, consider a single vector $\mathbf{x}$ and understand how to sample to preserve norm of $\mathbf{y} = \mathbf{A}\mathbf{x}$:

$$\|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 = \|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{S}\mathbf{y}\|_2^2 \approx \|\mathbf{y}\|_2^2 = \|\mathbf{A}\mathbf{x}\|_2^2.$$

Then we can union bound over an ϵ-net to extend to all $\mathbf{x}$.

VECTOR SAMPLING

As discussed a few lectures ago, uniform sampling only works well if $\mathbf{y} = \mathbf{A}\mathbf{x}$ is "flat".



Instead consider sampling with probabilities at least proportional to the magnitude of $\mathbf{y}$'s entries:

$$p_i > c \cdot \frac{y_i^2}{\|y\|_2^2} \text{ for constant } c \text{ to be determined.}$$

VARIANCE ANALYSIS

Let $\tilde{\mathbf{y}}$ be the subsampled $\mathbf{y}$. Recall that, when sampling with probabilities $p_1, \ldots, p_n$, for $i = 1, \ldots, n$ we add y_i to $\tilde{\mathbf{y}}$ with probability p_i and reweight by $\frac{1}{\sqrt{p_i}}$.

$$\|\tilde{\mathbf{y}}\|_2^2 = \sum_{i=1}^{n} \frac{y_i^2}{p_i} \cdot Z_i \quad \text{where} \quad Z_i = \begin{cases} 1 \text{ with probability } p_i \\ 0 \text{ otherwise} \end{cases}$$

$$\mathrm{Var}[\|\tilde{\mathbf{y}}\|_2^2] = \sum_{i=1}^{n} \frac{y_i^2}{p_i} \cdot \mathrm{Var}[Z_i] \leq \sum_{i=1}^{n} \frac{y_i^4}{p_i^2} \cdot p_i = \frac{y_i^4}{p_i}$$

We set $p_i = c \cdot \frac{y_i^2}{\|y\|_2^2}$ so get total variance:

$$\frac{1}{c}\|y\|_2^4$$

VARIANCE ANALYSIS

Using a Bernstein bound (or Chebyshev's inequality if you don't care about the δ dependence) we have that if $c = \frac{\log(1/\delta)}{\epsilon^2}$ then:

$$\Pr\left[\left|\|\tilde{\mathbf{y}}\|_2^2 - \|\mathbf{y}\|_2^2\right| \geq \epsilon\|\mathbf{y}\|_2^2\right] \leq \delta.$$

The number of samples we take in expectation is:

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} c \cdot \frac{y_i^2}{\|y_i\|_2^2} = \frac{\log(1/\delta)}{\epsilon^2}.$$

MAJOR CAVEAT!

We don't know $y_1, \ldots, y_n$! And in fact, these values aren't fixed. We wanted to prove a bound for $\mathbf{y} = \mathbf{A}\mathbf{x}$ for any $\mathbf{x}$.

Idea behind leverage scores: Sample row i from $\mathbf{A}$ using the worst case (largest necessary) sampling probability:

$$\tau_i = \max_{\mathbf{x}} \frac{y_i^2}{\|\mathbf{y}\|_2^2} \qquad \text{where} \qquad \mathbf{y} = \mathbf{A}\mathbf{x}.$$

If we sample with probability $p_i = \frac{1}{\epsilon^2} \cdot \tau_i$, then we will be sampling by at least $\frac{1}{\epsilon^2} \cdot \frac{y_i^2}{\|\mathbf{y}\|_2^2}$, no matter what $\mathbf{y}$ is.

Two concerns:

1) How to compute $\tau_1, \ldots, \tau_n$?

2) the number of samples we take will be roughly $\sum_{i=1}^{n} \tau_i$. How do we bound this?

LEVERAGE SCORE SAMPLING

Claim: $\tau_i = \|\mathbf{U}_i\|_2^2 = \max_x \frac{(\mathbf{A}x)_i^2}{\|\mathbf{A}x\|_2^2}$ is the i-th leverage score!

$$\begin{aligned}\frac{(\mathbf{A}x)_i^2}{\|\mathbf{A}x\|_2^2} &= \frac{(\mathbf{U}(\mathbf{\Sigma}\mathbf{V}^T x))_i^2}{\|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T x\|_2^2} \\ &= \frac{(\mathbf{U}(\mathbf{\Sigma}\mathbf{V}^T x))_i^2}{\|\mathbf{\Sigma}\mathbf{V}^T x\|_2^2} \\ &= \frac{(\mathbf{U}z)_i^2}{\|z\|_2^2} = \frac{\langle \mathbf{U}_i, z\rangle^2}{\|z\|_2^2} \leq \|\mathbf{U}_i\|_2^2\end{aligned}$$

where $z = \mathbf{\Sigma}\mathbf{V}^T x$. Here we used Cauchy-Schwarz's inequality:

$$\langle \mathbf{U}_i, z\rangle^2 \leq \|\mathbf{U}_i\|_2^2 \|z\|_2^2$$

Where equality holds when z is parallel to $\mathbf{U}_i$.

LEVERAGE SCORE SAMPLING

Leverage score sampling:

- For $i = 1, \ldots, n$,
  - Compute $\tau_i = \|\mathbf{U}_i\|_2^2$, where $\mathbf{U} \in \mathbb{R}^{n \times k}$ are left singular vectors of $\mathbf{A}$.
  - Set $p_i = \frac{c \log(1/\delta)}{\epsilon^2} \cdot \tau_i$.
  - Add row $\mathbf{a}_i$ to $\tilde{\mathbf{A}}$ with probability p_i and reweight by $\frac{1}{\sqrt{p_i}}$.

For any fixed $\mathbf{x}$, we will have that $(1-\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \leq \|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2$ with probability $(1-\delta)$.

How many rows do we sample in expectation?

SUM OF LEVERAGE SCORES

Claim: No matter how large $n \geq d$ is, for any matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, we have $\sum_{i=1}^{n} \tau_i \leq d$.

"Zero-sum" law for the importance of matrix rows.

LEVERAGE SCORE SAMPLING

Leverage score sampling:

- For $i = 1, \ldots, n$,
  - Compute $\tau_i = \|\mathbf{U}_i\|_2^2$, where $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.
  - Set $p_i = \frac{c \log(1/\delta)}{\epsilon^2} \cdot \tau_i$.
  - Add row $\mathbf{a}_i$ to $\tilde{\mathbf{A}}$ with probability p_i and reweight by $\frac{1}{\sqrt{p_i}}$.

For any fixed $\mathbf{x}$, we will have that
$(1-\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \leq \|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2$ with prob. $1-\delta$.

Since $\sum_{i=1}^n p_i = \frac{c \log(1/\delta)}{\epsilon^2} \cdot \sum_{i=1}^n \tau_i$, the sampled matrix $\tilde{\mathbf{A}}$ contains $O\left(\frac{d \log(1/\delta)}{\epsilon^2}\right)$ rows in expectation.

Last step: extend to all $\mathbf{x}$ with an ϵ-net!

MAIN RESULT

Naive ϵ-net argument leads to d^2 dependence since we need to set $\delta = c^d$. Gives "weaker" theorem:

Theorem (Subspace Embedding from Subsampling)

For each i, and fixed constant c, let $p_i = \min\left(1, \frac{cd}{\epsilon^2} \cdot \tau_i\right)$. Let $\tilde{\mathbf{A}}$ have rows sampled from $\mathbf{A} \in \mathbb{R}^{n \times d}$ with probabilities $p_1, \ldots, p_n$. With probability $9/10$, for all $x \in \mathbb{R}^d$:

$$(1-\epsilon)\|\mathbf{Ax}\|_2^2 \leq \|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{Ax}\|_2^2,$$

and $\tilde{\mathbf{A}}$ has $O(d^2/\epsilon^2)$ rows in expectation.

MAIN RESULT

Naive ϵ-net argument leads to d^2 dependence since we need to set $\delta = c^d$. Gives "weaker" theorem:

Theorem (Subspace Embedding from Subsampling)

For each i, and fixed constant c, let $p_i = \min\left(1, \frac{cd}{\epsilon^2} \cdot \tau_i\right)$. Let $\tilde{\mathbf{A}}$ have rows sampled from $\mathbf{A} \in \mathbb{R}^{n \times d}$ with probabilities $p_1, \ldots, p_n$. With probability $9/10$, for all $x \in \mathbb{R}^d$:

$$(1-\epsilon)\|\mathbf{Ax}\|_2^2 \leq \|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{Ax}\|_2^2,$$

and $\tilde{\mathbf{A}}$ has $O(d^2/\epsilon^2)$ rows in expectation.

Not good enough for graph sparsification!

If $G = (V, E)$, then $d = |V|$ and $n = |E|$, so $d^2 \geq n$, and we sample all edges!

IMPROVING TO $\tilde{O}(D/\epsilon^2)$ SAMPLES

Lets modify algorithm to sample only (and exactly) $k = O(\frac{d \log d}{\epsilon^2})$ rows of $\mathbf{A}$. Let $(q_1, \ldots, q_n)$ be the distribution over $[n]$ given by $q_i = \frac{\tau_i}{\sum_j \tau_j} = \frac{\|\mathbf{U}_i\|_2^2}{d}$.

- For $i = 1, \ldots, k$,
  - Sample $j \sim [n]$ from the distribution $(q_1, \ldots, q_n)$.
  - Add row $\mathbf{a}_j$ to $\tilde{\mathbf{A}}$ and reweight by $\frac{1}{\sqrt{kq_j}}$.

We can let $\mathbf{S} \in \mathbb{R}^{k \times n}$ be the sampling and re-scaling matrix, such that $\mathbf{SA} = \tilde{\mathbf{A}}$

Getting the improved $d \log d$ dependence requires a new tool: the Matrix Chernoff bound

MAIN RESULT

Theorem (Subspace Embedding from Subsampling)

For each i, and fixed constant c, let $q = (q_1, \ldots, q_n)$ be the distribution over $[n]$ given by $q_i = \frac{\tau_i}{\sum_j \tau_j}$. Let $\mathbf{S} \in \mathbb{R}^{k \times n}$ be a row sampling matrix, where $k = O(\frac{d \log d}{\epsilon^2})$, such that $\mathbf{S}_i = \frac{1}{\sqrt{kq_j}} \cdot e_j$ for each row $i \in [k]$, where $j \sim_q [n]$ is drawn from the distribution q. With probability $9/10$, for all $x \in \mathbb{R}^d$:

$$(1-\epsilon)\|\mathbf{Ax}\|_2^2 \leq \|\mathbf{SAx}\|_2^2 \leq (1+\epsilon)\|\mathbf{Ax}\|_2^2,$$

and $\tilde{\mathbf{A}}$ has $O(d \log d/\epsilon^2)$ rows in expectation.

Goal: Prove this stronger theorem

SIMPLIFIED STRUCTURE

Claim 1: We can assume $\mathbf{A} = \mathbf{U} \in \mathbb{R}^{n \times d}$ has orthogonal columns.

Proof: Convince yourself that the following two statements are equivalent:

- For all $x \in \mathbb{R}^d$: $\quad \|\mathbf{SA}x\|_2 = (1 \pm \epsilon)\|\mathbf{A}x\|_2$
- For all $x \in \mathbb{R}^d$: $\quad \|\mathbf{SU}x\|_2 = (1 \pm \epsilon)\|\mathbf{U}x\|_2$

In both cases, $\mathbf{A}x \in \mathbb{R}^n$ and $\mathbf{U}x \in \mathbb{R}^n$ range over the full k-dimensional subspace W spanned by the columns of $\mathbf{A}$! Equiv to

$$\|\mathbf{S}y\|_2 = (1 \pm \epsilon)\|y\|_2, \quad \text{for all } y \in W$$

SIMPLIFIED STRUCTURE

Claim 1: We can assume $\mathbf{A} = \mathbf{U} \in \mathbb{R}^{n \times d}$ has orthogonal columns.

Proof: Convince yourself that the following two statements are equivalent:

- For all $x \in \mathbb{R}^d$: $\|\mathbf{SA}x\|_2 = (1 \pm \epsilon)\|\mathbf{A}x\|_2$
- For all $x \in \mathbb{R}^d$: $\|\mathbf{SU}x\|_2 = (1 \pm \epsilon)\|\mathbf{U}x\|_2$

In both cases, $\mathbf{A}x \in \mathbb{R}^n$ and $\mathbf{U}x \in \mathbb{R}^n$ range over the full k-dimensional subspace W spanned by the columns of $\mathbf{A}$! Equiv to

$$\|\mathbf{S}y\|_2 = (1 \pm \epsilon)\|y\|_2, \quad \text{for all } y \in W$$

Thus, our goal for a subspace embedding is to show that $\|\mathbf{SU}x\|_2^2 = (1 \pm \epsilon)\|\mathbf{U}x\|_2^2$ for all $x \in \mathbb{R}^d$.

SIMPLIFIED STRUCTURE

Claim 2: It suffices to show that $\|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}\|_2 \leq \epsilon$.

Proof:

$$\begin{aligned} \left|\|\mathbf{S}\mathbf{U}x\|_2^2 - \|\mathbf{U}x\|_2^2\right| &= \left|x^T\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}x - x^T\mathbf{I}x\right| \\ &= \left|x^T\left(\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}\right)x\right| \\ &\leq \|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}\|_2\|x\|_2 \leq \epsilon\|\mathbf{U}x\|_2^2 \end{aligned}$$

Where we used $\|\mathbf{A}\|_2 = \max_x \frac{x^T\mathbf{A}x}{\|x\|_2}$ for any symmetric matrix $\mathbf{A}$.
Follows that

$$\|\mathbf{S}\mathbf{U}x\|_2^2 = (1 \pm \epsilon)\|\mathbf{U}x\|_2^2$$

LEVERAGE SCORE SAMPLING

Recall our algorithm, that samples $k = O(\frac{d \log d}{\epsilon^2})$ rows of $\mathbf{A}$. Let $(q_1, \ldots, q_n)$ be the distribution over $[n]$ given by $q_i = \frac{\tau_i}{\sum_j \tau_j} = \frac{\|\mathbf{U}_i\|_2^2}{d}$.

- For $i = 1, \ldots, k$,
  - Sample $j \sim [n]$ from the distribution $(q_1, \ldots, q_n)$.
  - Add row $\mathbf{a}_j$ to $\tilde{\mathbf{A}}$ and reweight by $\frac{1}{\sqrt{kq_j}}$.

We can let $\mathbf{S} \in \mathbb{R}^{k \times n}$ be the sampling and re-scaling matrix, such that $\mathbf{SA} = \tilde{\mathbf{A}}$

Summary of Claims 1 + 2: It suffices to show that

$$\|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}\|_2 \leq \epsilon$$

where $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD.

SIMPLIFIED STRUCTURE

Summary: It suffices to show that $\|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}_d\|_2 \leq \epsilon$, where $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD.

Let $s_j \in [n]$ be the index of the j-th row we sample in our algorithm. We have

$$\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} = \sum_{j=1}^{k} \frac{\mathbf{U}_{s_j}^T\mathbf{U}_{s_j}}{k \cdot q_{s_j}}$$

Notice that

$$\mathbb{E}_{s_j}\left[\sum_{j=1}^{k} \frac{\mathbf{U}_{s_j}^T\mathbf{U}_{s_j}}{k \cdot q_{s_j}}\right] = \sum_{j=1}^{k}\sum_{i=1}^{n} q_i \cdot \frac{\mathbf{U}_i^T\mathbf{U}_i}{kq_i}$$
$$= k \cdot \frac{1}{k}\mathbf{U}^T\mathbf{U} = \mathbf{I}_d$$

SIMPLIFIED STRUCTURE

We have $\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} = \sum_j \frac{\mathbf{U}_{s_j}^T\mathbf{U}_{s_j}}{kq_{s_j}}$, and

$$\frac{1}{k}\mathbb{E}_{s_j}\left[\sum_{j=1}^{k} \frac{\mathbf{U}_{s_j}^T\mathbf{U}_{s_j}}{q_{s_j}}\right] = \mathbf{I}_d$$

If we define $\mathbf{X}_j = \mathbf{I}_d - \frac{\mathbf{U}_{s_j}^T\mathbf{U}_{s_j}}{q_{s_j}}$, we have $\mathbb{E}[\frac{1}{k}\sum_{j=1}^k \mathbf{X}_j] = 0$, and

$$\frac{1}{k}\sum_{j=1}^{k}\mathbf{X}_j = \mathbf{I}_d - \mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}$$

Now we want concentration: show $\frac{1}{k}\sum_{j=1}^k \mathbf{X}_j$ is close to its expectation!

RANDOM MATRIX CONCENTRATION

We have i.i.d. random matrices $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$, with mean zero: $\mathbb{E}[\frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i] = 0$.

We need $\frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i$ to concentrate around its expectation in *spectral norm!*

We want:

$$\left\|\frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i\right\|_2 = \|\mathbf{I}_d - \mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}\|_2 < \epsilon$$

with high probability.

To achieve this, we will use *Matrix Concentration Inequalities*!

MATRIX CHERNOFF BOUNDS

Generalization from concentration of sums of random numbers, to sums of random matrices.

Theorem

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ *be i.i.d. copies of a symmetric random matrix* $\mathbf{X} \in \mathbb{R}^{d \times d}$*, with*

- $\mathbb{E}[\mathbf{X}] = 0$ *(zero mean)*
- $\|\mathbf{X}\|_2 \leq \gamma$ *with probability* 1*. (bounded norm)*
- $\|\mathbb{E}[\mathbf{X}^T\mathbf{X}]\|_2 \leq \sigma^2$*. (matrix variance)*

Then for any $\epsilon > 0$*, we have*

$$\Pr\left[\left\|\frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i\right\| > \epsilon\right] \leq 2d \cdot e^{-\frac{k\epsilon^2}{\sigma^2 + \gamma\epsilon/3}}$$

MATRIX CHERNOFF BOUNDS

Recall: $\mathbf{X} = \mathbf{I}_d - \frac{\mathbf{U}_i^T \mathbf{U}_i}{q_i}$, where $i \sim [n]$ is sampled according to $(p_1, \ldots, p_n)$. We have $\mathbb{E}[\mathbf{X}] = 0$.

$$\|\mathbf{X}\|_2 \leq \|\mathbf{I}_d\|_2 + \max_i \|\frac{\mathbf{U}_i^T \mathbf{U}_i}{q_i}\|_2 \leq 1 + \max_i \frac{\|\mathbf{U}_i\|_2^2}{q_i} \leq 1 + d$$

MATRIX CHERNOFF BOUNDS

Recall: $\mathbf{X} = \mathbf{I}_d - \frac{\mathbf{U}_i^T \mathbf{U}_i}{q_i}$, where $i \sim [n]$ is sampled according to $(p_1, \ldots, p_n)$. We have $\mathbb{E}[\mathbf{X}] = 0$.

$$\|\mathbf{X}\|_2 \leq \|\mathbf{I}_d\|_2 + \max_i \|\frac{\mathbf{U}_i^T \mathbf{U}_i}{q_i}\|_2 \leq 1 + \max_i \frac{\|\mathbf{U}_i\|_2^2}{q_i} \leq 1 + d$$

$$\begin{aligned}\left\|\mathbb{E}\left[\mathbf{X}^T\mathbf{X}\right]\right\|_2 &\leq \mathbf{I}_d - 2\mathbb{E}_{i\sim(q_1,\ldots,q_n)}\left[\frac{\mathbf{U}_i^T\mathbf{U}_i}{q_i}\right] + \mathbb{E}_{i\sim(q_1,\ldots,q_n)}\left[\frac{\mathbf{U}_i^T\mathbf{U}_i\mathbf{U}_i^T\mathbf{U}_i}{q_i^2}\right] \\ &= \sum_i \frac{\mathbf{U}_i^T\mathbf{U}_i\mathbf{U}_i^T\mathbf{U}_i}{q_i^2} \cdot q_i - \mathbf{I}_d \\ &\preceq d\sum_i \mathbf{U}_i^T\mathbf{U}_i - \mathbf{I}_d \preceq (d-1)\mathbf{I}_d\end{aligned}$$

So $\gamma < O(d)$ and $\sigma^2 \leq O(d)$.

MATRIX CHERNOFF BOUNDS

Theorem

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ be i.i.d. copies of a symmetric random matrix $\mathbf{X} \in \mathbb{R}^{d \times d}$, with $\mathbb{E}[\mathbf{X}] = 0$, $\|\mathbf{X}\|_2 \leq \gamma$, and $\|\mathbb{E}[\mathbf{X}^T \mathbf{X}]\|_2 \leq \sigma^2$. Then for any $\epsilon > 0$, we have

$$\Pr\left[\left\|\frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i\right\| > \epsilon\right] \leq 2d \cdot e^{-\frac{k\epsilon^2}{\sigma^2 + \gamma\epsilon/3}}$$

We have $\gamma < O(d)$ and $\sigma^2 \leq O(d)$. So setting $k = (d \log d/\epsilon^2)$

$$\Pr\left[\left\|\mathbf{U}^T\mathbf{U} - \mathbf{I}_d\right\|_2 > \epsilon\right] \leq 2d \cdot e^{-\frac{k\epsilon^2}{\Theta(d)}} \leq \frac{1}{d}$$

This is what we needed to show!

MAIN RESULT

Using matrix concentration inequalities, we obtain the tighter bound of $k = O(\frac{d \log d}{\epsilon^2})$ samples.

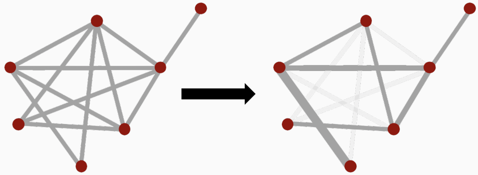
Theorem (Subspace Embedding from Subsampling)

For each i*, let* $q_i = \frac{\|\mathbf{U}_i\|_2^2}{d}$*. Let* $\tilde{\mathbf{A}} \in \mathbb{R}^{k \times n}$ *have* $k = O(\frac{d \log d}{\epsilon^2})$ *rows sampled from* $\mathbf{A} \in \mathbb{R}^{n \times d}$ *via the distribution* $(q_1, \ldots, q_n)$*, and scaled by* $1/\sqrt{q_i k}$*. With probability* $9/10$*, for all* $x \in \mathbb{R}^d$*:*

$$(1-\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \leq \|\tilde{\mathbf{A}}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}\|_2^2,$$

SPECTRAL SPARSIFICATION COROLLARY

For any graph G with n nodes with m edges, there exists a graph $\tilde{G}$ with $O(n \log n/\epsilon^2)$ edges such that, for all $\mathbf{x}$, $\|\tilde{\mathbf{B}}\mathbf{x}\|_2^2 = (1 \pm \epsilon)\|\mathbf{B}\mathbf{x}\|_2^2$.



As a result, the value of any cut in $\tilde{G}$ is within a $(1 \pm \epsilon)$ factor of the value in G, the Laplacian eigenvalues are with a $(1 \pm \epsilon)$ factors, etc.

FAST ALGORITHMS FOR MAX FLOW/MIN CUT

Theorem: There is an algorithm for computing a $(1-\epsilon)$ optimal max s-t flow in time $O(mn^{1/3}\,\mathrm{poly}(1/\epsilon))$, and a min s-t cut in time $O(m+n^{4/3}\,\mathrm{poly}(1/\epsilon))$.

Electrical flows, Laplacian systems, and faster approximation of maximum flow in undirected graphs, Christiano, Kelner, Madry, Spielman, Teng (STOC '11)

Rough idea: Sparisfy graph, then run known max-flow/min-cut algorithms on spectral sparsifier.